
so AI is for speed. How would using slower and less heat generating processors work. Clients would have wait longer for answers. What if the answers came back in 3 minutes to several hours?

Leon Maxey <suleonmax@gmail.com>
Draft

Sat, Jun 27, 2026 at 12:34 PM

Using slower, less power-hungry processors that delay answers by minutes or hours would fundamentally change how AI infrastructure works. This approach is called **asynchronous or batch processing**, and it is highly effective for reducing energy and water consumption. [1, 2, 3]

By removing the requirement for instant, split-second answers, tech companies can run data centers in a completely different way.

1. Servers Can Use "Underclocking" to Run Cooler

When a computer processor is pushed to operate at maximum speed, its heat generation and power consumption skyrocket exponentially. [4]

- **The Physics:** Slowing a processor down by just 20% can reduce its heat generation and power usage by up to 50%.
- **The Impact:** If an AI model is allowed 10 minutes to think instead of two seconds, the servers can be "underclocked" to run at a slow, gentle pace. They stay naturally cool, completely eliminating the need for massive, evaporating water-cooling towers. [5]

2. AI Shifts to Nighttime and Winter Schedules

Right now, data centers must run at full blast during the hot afternoon when millions of people are active online, which requires maximum cooling.

- **The Impact:** If users tolerate a multi-hour delay, companies can queue up non-urgent requests and run them strictly at night when the outside air is cold enough to cool the servers naturally without using water. They can also schedule heavy processing to occur only when there is excess, cheap solar or wind energy available on the power grid.

3. Smaller, Specialized Microchips Become Viable

Instead of relying on massive, power-hungry Graphics Processing Units (GPUs) that pull massive amounts of electricity, a slower system could use small, hyper-efficient microchips.

- **The Impact:** These smaller chips take longer to process data, but they can be cooled passively using simple metal heat sinks and small fans—the same way a standard home laptop or stereo amplifier stays cool without liquid.

Where This Delay Works (and Where It Fails)

This trade-off is already used for certain tasks, but it does not work for everything. [6]

- **Where a 3-hour delay works perfectly:** Tasks like drafting a long-form essay, translating a book into another language, analyzing weekly business spreadsheets, or generating complex digital art. A user can submit the prompt, walk away, and check the results later. [7]
- **Where a 3-hour delay fails:** Real-time tasks like searching for breaking news, using voice assistants (like Siri or Alexa), live customer service chatbots, or coding assistants that programmers use line-by-line while working. [8]

If you want, I can:

Explain how **batch processing** currently works for industrial AI tasksShow you how **neuromorphic chips** mimic the human brain to use less power

- [1] <https://www.sciencedirect.com>
- [2] <https://hookdeck.com>
- [3] <https://blog.everpuredata.com>
- [4] <https://www.reddit.com>
- [5] <https://www.instagram.com>
- [6] <https://openreview.net>
- [7] <https://medium.com>
- [8] <https://www.sandgarden.com>